Learning from Pathology Databases to Improve the Laboratory Diagnosis of Infectious Diseases

Alice Richardson[1], Fariba Shadabi[1], Brett A. Lidbury[2]

[1] Faculty of Information Sciences and Engineering
University of Canberra, ACT 2601, Australia
[2] Centre for Biomedical and Forensic Research, Faculty of Applied Science
University of Canberra, ACT 2601, Australia

Abstract. *This paper investigates the effect of data pre-processing and the use of ensemble on the accuracy of decision trees. The methodology is illustrated using a previously unanalysed data set from ACT Pathology (Canberra, Australia) relating to Hepatitis B and Hepatitis C patients.*

Keywords: Decision Tree, Ensemble Classifier, Data Transformation, Hepatitis.

1 INTRODUCTION

In this paper we describe our empirical study of constructing a decision tree ensemble using different data pre-processing techniques on multi-variable pathology laboratory data for the enhanced laboratory diagnosis of infectious diseases (for this study, hepatitis B and hepatitis C viruses). We specifically use a data set of 18625 deidentified records from 1997 – 2007 made available to us by ACT Pathology. Seventeen explanatory variables and two response variables were provided.

2 METHODOLOGY AND EXPERIMENTAL RESULTS

In this study we employed S-PLUS decision trees and carried out a four-factor experiment to study the effect of virus, outcome, method and preprocessing on the overall accuracy rate. There are two viruses (Hepatitis B (HBV) and Hepatitis C (HepC)) and two outcomes (positive or negative). There are three methods (basic multiple, majority multiple and clear negative).
The basic multiple method consists of a standard decision tree with 2/3 of the data for training and 1/3 for testing. The 2/3 of negative outcomes are split into 72 sets, each the same size as the 2/3 of positive cases. Accuracy rate is calculated for each tree separately.

The majority multiple method uses 36 subsets for training where each subset had 282 negative outcomes and 141 positive outcomes for HBV. Furthermore, we

computed the accuracy rate for each tree (using the same test dataset) based on majority voting from all trees.

The clear negative method involves selecting the cases which are "certainly negative", this gave us total of 154 cases. We then combined this with 1/3 of the positive cases to construct the training set. The remaining data are used for testing.

Finally, there are four ways of preprocessing: none, scaling, logging and scale-logging. Scaling sets the range of each explanatory variable to a common range of 0 – 100. Logging uses the natural logarithm transformation. Scale-logging uses a common range of 0 – 100 then takes the natural logarithm. Note also that assignment of positive and negative to data occurs before scaling.

The analysis of variance shows that accuracy rate depends on outcome ($F = 32.279$, $df = 1$ and 23, $p = .000$). Positive cases have a higher accuracy rate on average than negative. There are also a significant interaction between method and outcome ($F = 50.640$, $df = 2$ and 23, $p = .000$). Majority multiple does better on average at predicting negatives, whereas the other two methods do better on average at predicting positives. The other significant interaction is between virus and outcome ($F = 32.120$, $df = 1$ and 23, $p = .000$). For HepC positive leads to higher accuracy rates on average, the reverse is true for HBV.

3 DISCUSSION AND CONCLUSION

Immunoassay techniques are routinely used in pathology departments to detect antibodies to disease-causing microbes, indicating previous exposure to, or infection by, the specific pathogen. This study examined the immunoassay marker HBSA to detect previous hepatitis B virus (HBV) infection/exposure, and the general HepC antigen to detect hepatitis C virus (HCV) infection.

The best approach for negative HBSA and HepC data accuracy was the "basic single" method due to the size of the dataset.

For smaller datasets, as found for both HBSA and HepC positive cohorts, other methods were required to achieve high predictive accuracy based on associated pathology data (described in Table1). Furthermore, the "clear negative" method, which used other pathology data (i.e. ALT liver enzyme) to give the most certain true negative cohort, was very effective. For this method, patient data with HBSA ≤ 0.01 and ALT ≤ 55 U/L were considered to be "clear negative" for HBV. We also considered patient data with HepC <= 0.03 as "clear negative" for HCV.

Acknowledgements:

The authors wish to thank Mr Gus Koerbin, Principal Scientist, ACT Pathology, and his staff, for their support of and interest in this study. The authors also thank "Medical Advances Without Animals (MAWA)" for funding this project.